DOCUMENT RESUME

ED 174 638 TM 009 236

AUTHOR Garrison, Wayne M.; White, Karl R.
TITLE A Simulation Study on the Utility of Rasch and Classical Test Analysis Procedures.
PUB DATE Apr 79
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Item Analysis; Models; Simulation; Test Construction; *Test Interpretation; Test Reliability
IDENTIFIERS Monte Carlo Method; *Rasch Model

ABSTRACT

Rasch and classical test analysis methods were compared with respect to their similarities and differences in the identification of noninformative items and implausible person records. Using computer simulated data with known parameters, each model was evaluated in terms of its effectiveness in: (1) identifying noninformative or "bad" items; (2) identifying persons for whom an instrument may be inappropriate; and (3) yielding an observed test score which correlated highly with a known true score. Results of the analyses demonstrated the superiority of the Rasch model over classical item analysis procedures, particularly in the identification of persons for whom an instrument fails to provide an accurate measure of the variable in question. (Author/JKS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED174638

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Simulation Study on the Utility of Rasch and Classical Test Analysis Procedures

Wayne M. Garrison

National Technical Institute for the Deaf

Rochester Institute of Technology

and

Karl R. White

Exceptional Child Center

Utah State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

W. Garrison

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

TM009 236

Abstract

The purpose of this study was to compare Rasch and classical test analyses with respect to their similarities and differences in the identification of noninformative items and implausible person records. Using computer simulated data with known parameters, each model was evaluated in terms of its effectiveness in: (1) identifying noninformative or "bad" items; (2) identifying persons for whom an instrument may be inappropriate; and (3) yielding an observed test score which correlated highly with a known true score. Results of the analyses demonstrated the superiority of the Rasch model over classical item analysis procedures, particularly in the identification of persons for whom an instrument fails to provide an accurate measure of the variable in question.

A Simulation Study on the Utility of Rasch and Classical Test Analysis Procedures

In recent years, substantial interest has developed among testing specialists and practitioners in the development and applicability of mathematical models to educational and psychological assessment problems. In particular, the one-parameter Rasch logistic model advanced by Wright and his associates over the past decade (e.g., Wright, 1968; Wright, 1977; Wright & Mead, 1975, 1977; Wright & Stone, in press) has generated considerable enthusiasm in its application and solution to many measurement problems previously regarded as intractable.

The Rasch model, like a number of other measurement models, has been referred to as a latent trait model. That is, the model seeks to provide an explanation of an examinee's test performance in terms of the unobservable traits, abilities, emotions or inclinations which presumably underlie performance on a given test. The Rasch model also differs from other latent trait models in its inclusion of the fewest parameters. Thus, the model postulates that the outcome of the interaction between persons and items in a testing situation is explainable simply as a function of the difference between an examinee's ability or position on some latent continuum and the difficulty of items scaled along the same continuum.

Proponents of the Rasch model claim that it offers significant advantages over classical approaches to test analysis. The primary advantage claimed is the capability of estimating both item and person

parameters which are relatively invariant across different subsets of items and sub-groups of persons. By placing both item and person parameters on a common metric, Rasch analyses also purportedly produce a degree of measurement precision not previously attainable within a classical test theory framework.

While the Rasch model has now been studied and applied in a variety of measurement situations, it is, nonetheless, surprising that few studies have investigated the model in terms of its predictive validity. The model has been studied for robustness (e.g., Hambleton & Traub, 1976). Calibration of tests through Rasch analyses have been found to improve instrument internal consistency through the elimination of misfitting items (Ryan & Hamm, 1976). The model also has been used to equate different tests of the same latent trait to a common scale (Rentz & Bashaw, 1977). Yet, the predictive validity of its results in applications to test data has not been explored fully, likely attributable to the problem of obtaining appropriate criteria. Coupled with this situation, there remain many questions among measurement practitioners as to the effectiveness of the Rasch model relative to classical item analysis procedures.

In practice, educational and psychological tests which utilize an objective response format may be subjected to a number of different item analysis procedures. Since there are significant operational and theoretical differences among these procedures, it is not unreasonable to suppose that two test developers calibrating a common pool of items via different item analytic methods may arrive at differing conclusions regarding the item composition of the "best" test. Each, presumably, could argue that

his/her test was the best (i.e., most reliable or valid) measure of the construct in question. However, since the true scores of individuals on the construct being tested are almost never known, it is difficult to determine which item analysis procedure results in the best version of a given test.

This study represented an attempt to address some of the questions surrounding the applications of Rasch and classical test analysis procedures to test analysis problems. Specifically, the purpose of the investigation was to compare Rasch and classical test analysis procedures with respect to the effectiveness of each method in: (1) identifying items which are not an accurate measure of the construct in question; (2) identifying persons for whom an instrument is inappropriate; and (3) yielding a total observed test score which is correlated highly with a "known" true score. All data were simulated with known characteristics.

Method

Procedure

Using a computer simulation program (DATASIM: Baidoo, 1978), test scores on an hypothetical construct (Y) were generated for 2,500 persons. These scores were normally distributed with a mean of 50 and a standard deviation of 10 (permissible range = 100). From this distribution of 2,500 data points, representing the distribution of true scores (Y_n) for the population, a sample of 200 persons was randomly selected for subsequent analyses.

An observed total test score (y_n) for each of these 200 persons was determined as follows: For each Y_n, a normal distribution of 500 data points was generated with a mean of Y_n and a standard deviation of 3.

The standard deviation of this distribution represents the standard error of measurement (S_e) for the test. One of these 500 scores was then randomly selected as the individual's observed score (y_n).

Using the classical definition of reliability (r_{tt}) as the ratio of the true score variance (S_Y^2) to the observed score variance (S_y^2), it can be shown that $S_e = S_y \sqrt{1 - r_{tt}}$. Substituting the known parameters in this equation, a group of test scores with a standard error of measurement of 3 will have a theoretical reliability of .91. Moreover, since the true (Y_n) and the observed (y_n) scores for each individual are known, the validity of the test can be estimated by the correlation between true and observed scores (r_{Yy}).

The true scores (Y_n) for each person were ranked from highest to lowest and the items were ranked from easiest to hardest. Consequently, a person with a true score of 57 answered the first 57 items correctly and answered the last 43 items incorrectly. The true score persons x items matrix formed represented a perfect Guttman scale including 100 items and 200 persons.

The observed score persons x items matrix used in the analyses was then formed by examining the difference between the total true and observed scores ($Y_n - y_n$) for each person. If, for the n^{th} person, the true score was 57 and the observed score was 51, six of the first 57 items were randomly selected and correct responses on these items were changed to incorrect responses. Conversely, a person with a true score of 57 and an observed score of 61 would have 4 of the last 43 items (i.e., items 58-100) randomly selected and these incorrect responses changed to correct responses. The true score item response vectors for

each person were then adjusted to form the observed score persons x items data matrix. All 100 of the items included in the test represented "good" items (i.e., items which were a valid measure of the construct in question). As is found in practice, however, a pre-specified amount of error variance was built into each item.

Two other conditions were necessary to enable the critical comparison of the two item analytic procedures. First, 25 "bad" items were added to the data set for each of the 200 persons. The responses for these 25 items consisted of random sequences of 1's and 0's (denoting correct and incorrect responses, respectively). Such a response pattern would be expected for items which were ambiguous, miskeyed, or possibly measuring some irrelevant construct. This resulted in a total test consisting of 125 items. Secondly, 25 persons who had random responses across all 125 items were added to the data matrix. This type of response pattern would be expected for persons who, for some reason, were not performing to their capabilities, failed to understand the directions or language of the test, etc. (Such people are referred to as "misfitting" people throughout this paper). For such a person, the observed score on the test (even if it were a good test) would not be an accurate measure of the construct in question. The true scores for these misfitting persons were randomly selected from the same population as were the true scores for the 200 "good" persons ($\bar{X}$ = 50, SD = 10). The final observed score for each of the 225 people was the number of correct responses across all 125 items.

The final persons x items observed score data matrix is illustrated in Figure 1. As can be seen, the data set provides information about

the true score and observed score of persons for whom the test was an accurate assessment of the construct in question, and for those for whom the test was not an accurate assessment. In addition, it provides item information for all people on both good and bad items. The comparison of the two item analysis procedures was accomplished by examining: (1) the capability of each procedure to identify and exclude bad items; (2) the capability of the Rasch model to identify misfitting persons and the influence of such misfitting persons on the classical analysis test characteristics; and (3) the characteristics (i.e., reliability, validity, number of items, etc.) of the final test version which was derived using each of the two methods.

Insert Figure 1 here

Analyses

The simulated data set described above was subjected to both classical and Rasch test analyses. Classical test analyses were performed using program LERTAP (Nelson, 1974) and Rasch calibrations were performed using program BICAL (Wright & Mead, 1977).

The criterion used throughout the classical analyses to eliminate bad or noninformative items from the data matrix was that of item to total observed test score correlation. Consequently, items with point-biserial correlations (r_{pb}) less than .20 were rejected. This level was relaxed to .15 for items in the difficulty ranges .00 - .10 and .90 - 1.00. No justifiable provision for the elimination of person records exists for the classical item analysis approach.

Person and item eliminations for the Rasch analyses were determined using an algorithm provided by Wright and Stone (in press). This algorithm summarizes for both items and persons the degree to which a set of test responses corresponds to the expectations of the mathematical model and is expressed conveniently as a t-statistic. Persons with a t-statistic greater than 3.0 were eliminated from the test analyses as providing implausible test response vectors. A series of iterative item analyses utilized the same criterion for identifying misfitting or noninformative items.

Results and Discussion

Table 1 summarizes the characteristics of the original data set and the final versions of the tests derived using the different test analysis procedures. Although the final versions of the test are similar, there are a number of important differences. The most significant difference between the two procedures is the capability of the Rasch model to identify the misfitting persons included in the test calibrations.

Insert Table 1 here

The presence of these misfitting people in the classical item analysis resulted in a final version of the test which correlated less strongly with the true score, had a larger standard error of measurement and had slightly weaker internal consistency. Most of the differences between the two final test versions disappeared or were significantly reduced when the 25 misfitting persons were excluded by hand prior to the classical test analysis.

The impact of these misfitting people on the final test characteristics when a classical item analysis is applied is not surprising since the basis of the classical approach is the item to total observed score correlation. When 10% of the test population are responding in a random fashion, it is very likely that these coefficients will be adversely affected.

If a test developer is confident that there will not be a significant number of people in the tested population who will not attempt to do their best, against whom the test is biased, or who do not understand the directions of the test, then the Rasch model and the classical approach to test analysis and development are likely equally effective in developing a reliable and valid assessment instrument. Even in this situation, however, the Rasch model is slightly more efficient in that it achieves the same or better levels of reliability, validity and measurement error using fewer items.

The possibility that there exists a significant number of "misfitting" people in the tested population is not merely conjectural. It is plausible that many groups to which educational and psychological tests are administered will contain at least 10% of the study sample for whom a particular instrument is not an accurate measure of the construct in question. Whenever this situation exists, the Rasch model offers distinct advantages over classical approaches to item analysis. For this study, the classical approach to test analysis resulted is a correlation of .90 between true scores and observed scores on the final test version. On first observation, this is a respectable figure. However, the correlation between observed and true scores for the 25 misfitting people was only .18. Obviously,

any selection, treatment or placement decisions based on this test for these misfitting persons would be highly inappropriate and little better than decisions based upon no test at all. Although the Rasch model does not provide a way to estimate a better score for these persons, it is highly effective in selecting the persons for whom a particular test is inappropriate so that alternative testing strategies may be utilized.

To further understand how the "misfitting" persons negatively influence tne results of classical item analyses and test construction, and how the Rasch model and the classical methodology differ in their final selection of items, it is useful to consider the characteristics of some specific items. Figure 2(A) represents, for both a very easy and a very difficult item, the familiar item characteristic curve. Each of the curves illustrates the proportions of successes on the particular item as a function of ability level (total observed scores on the final version of the test).

Insert Figure 2 here

Both of the items illustrated in Figure 2(A) show a peculiar reversal phenomenon. That is, at fixed points along the ability continuum, it can be seen that more able respondents perform less well than do respondents of lower ability. This situation is identified partially by the r_{pb} coefficients computed for each of the items. Each of the items failed to achieve the $r_{pb} = .15$ criterion established for their difficulty levels and consequently, did not enter the final (refined) item pool.

Figure 2(B) represents the item response curves for the same two items resulting from the Rasch test analyses. However, the reversal is not present. Moreover, these same two items were identified as fitting the Rasch model and were included in the final collection of refined items.

The seemingly paradoxical situation described above is explainable in terms of the particular respondents utilized in each of the classical and Rasch calibration procedures. The classical analyses included all 225 respondents, whereas the Rasch procedure successfully identified and removed the random response patterns of the 25 misfitting persons prior to the calibration of items. What is reflected in Figure 2(A), then, is the negative influence on item properties of misfitting persons only. Removal of these persons from the classical item analyses would have resulted in retention of these items in the final item pool. However, the classical test model has no provision for such an elimination of implausible persons, resulting instead in the loss of otherwise useful items.

Figure 3 demonstrates another interesting difference between the final pool of items selected through the classical procedure (Panel A) and those selected by the Rasch analysis (Panel B). The difficulty indices on the x-axis are expressed as the familiar p (proportion correct) and in the standard score Δ scale used by the Educational Testing Service. The Δ values have the advantage of placing the difficulty indices on an interval scale with equal units. The discrimination indices r_{bis} are presented on the y-axis. The coefficient r_{bis} is used in this instance,

rather than r_{pb}, as a means of more fully differentiating indices of discriminating power and item difficulty.

Insert Figure 3 here

An examination of the information contained in Figure 3(A) indicates a concentration of items in the middle range of difficulty, with discrimination falling off sharply as items move away from the 50 percent level. This is precisely the representation one would predict from classical test theory.

In contrast to the above, Figure 4(B) shows a different relationship between the difficulty and discrimination parameters for items fitting the Rasch model. Here, the distribution of difficulty indices is essentially platykurtic, providing for more equal accuracy of measurement and maximum discrimination throughout a broader range of scores.

Conclusions

Given an initial pool of items, the goal of both the Rasch model and the classical approach to item analysis and test development is to identify the group of items which yields the "best" (i.e., most reliable and valid) measure of the construct in question. This study examined the relative effectiveness of each of these methods of test analysis in accomplishing this goal.

The results indicated that the Rasch model offers distinct advantages over classical item analysis procedures. The most important advantage which the Rasch model offers is the demonstrated capability of identifying

misfitting persons (i.e., those people against whom a test is biased, those who didn't understand the directions, or those who were not working up to their capabilities, etc.).

It is probable that many educational and psychological testing groups contain some number of such misfitting people. In the light of this situation, the results of the present study demonstrate two serious consequences of the inability of classical item analysis procedures to identify such persons. First, the observed test score for these persons, even on a relatively good set of items, will likely be inappropriate and misleading. Secondly, the presence of such misfitting persons in the tested population has a negative impact on the selection of items and the final characteristics of a test. The effectiveness of the Rasch model in identifying misfitting persons circumvents both of these problems. Even in those instances where there are no misfitting persons present in the population to be tested, the results of this study suggest that the Rasch model approach to test analysis is equal to or slightly better than the classical approach in terms of the efficiency (number of items included), reliability, validity and standard error of measurement of the test.

References

Baidoo, K. DATASIM: A procedure for the generation of random normal deviates. Unpublished manuscript, National Technical Institute for the Deaf, 1978.

Hambleton, R. K., & Traub, R. E. The robustness of the Rasch test model. Laboratory of Psychometric and Evaluative Research Report No. 42. Amherst, Mass.: University of Massachusetts, 1976.

Nelson, L. R. Guide to LERTAP use and interpretation. Dunedin, New Zealand: Department of Education, University of Otago, 1974.

Rentz, R. R., & Bashaw, W. L. The National Reference Scale for Reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14(2), 161-179.

Ryan, J. P., & Hamm, D. W. Practical procedures for increasing the reliability of classroom tests by using the Rasch model. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1976.

Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J.: Educational Testing Service, 1968, 85-101.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14(2), 97-116.

Wright, B. D., & Mead, R. J. CALFIT: Sample-free calibration with a Rasch measurement model. Research Memorandum No. 18, Statistical Laboratory, Department of Education, University of Chicago, 1975.

Wright, B. D., & Me l, R. J. BICAL: Calibrating items and scales with the Rasch model. Research Memorandum No. 23, Statistical Laboratory, Department of Education, University of Chicago, 1977.

Wright, B. D., & Stone, M. H. Best test design: A handbook for Rasch measurement. Palo Alto, Calif.: Scientific Press, in press.

DG2A&B:Garrison
r2/1/3/79:jo

Acknowledgements

This document was developed in the course of an agreement with the United States Department of Health, Education, and Welfare.

		Persons									j = 1,225		
		1	2	. . .		m	. . .		200		201	. . .	225
i = 1,125	1												
	2												
	.												
	.												
	.												
						Good Items Good Persons					Good Items Bad Persons		
	8												
	.												
Items	.												
	.												
	100												
	101												
	.												
	.					Bad Items Good Persons					Bad Items Bad Persons		
	.												
	125												
Observed Scores (y_n)		y_1	y_2	. . .		y_m	. . .		y_{200}		. . .		y_{225}

Figure 1. Simulated data matrix.

Table 1
Summary Findings from the Classical and
Rasch Test Analyses

TEST PROPERTIES	Original Data Set	After Rasch Analysis	After Classical Analysis	After Classical Analysis*
Number of items	125	40	49 (35)**	60
r_{tt} (Hoyt index)	.86	.96	.94	.96
s_m(Based on r_{tt})	4.07	1.89	2.39	2.21
r_{YY}(Validity)	.86	.99	.90	.99
Number of "bad" items excluded		25	25	25
Number of "good" items excluded		60	51	40
Number of "misfitting" people identified and excluded		25	0	25 (By Hand)
Number of "good" people identified and excluded		0	0	0

* Misfitting persons removed by hand prior to test analyses.
** Number of items in classical analysis common to the Rasch analysis.

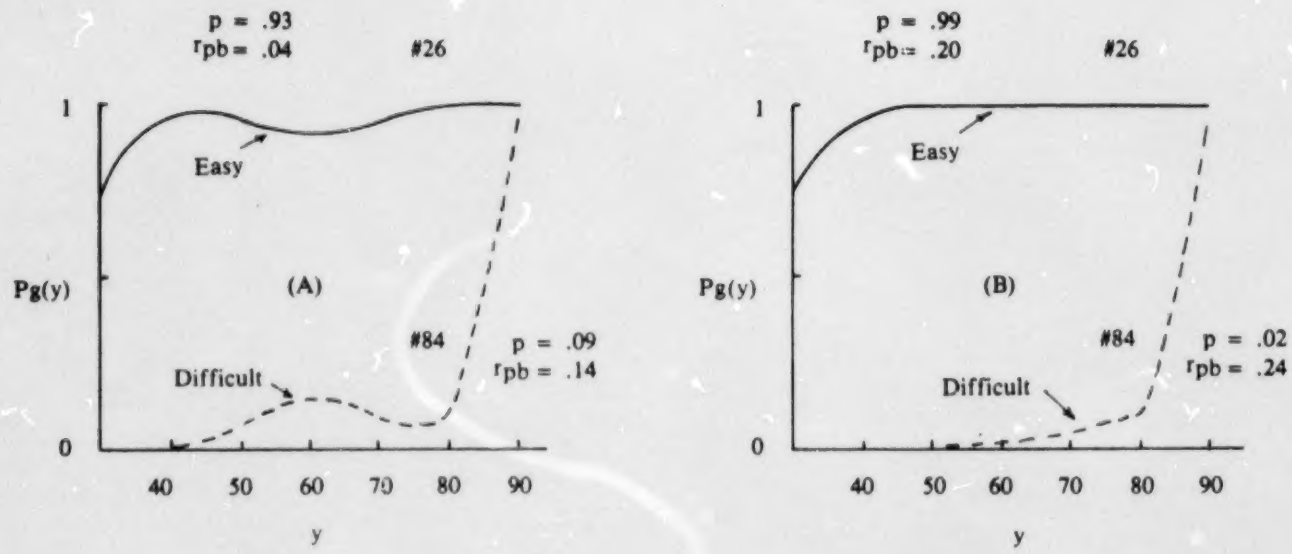

Figure 2. Item response curves for items 26 and 84.

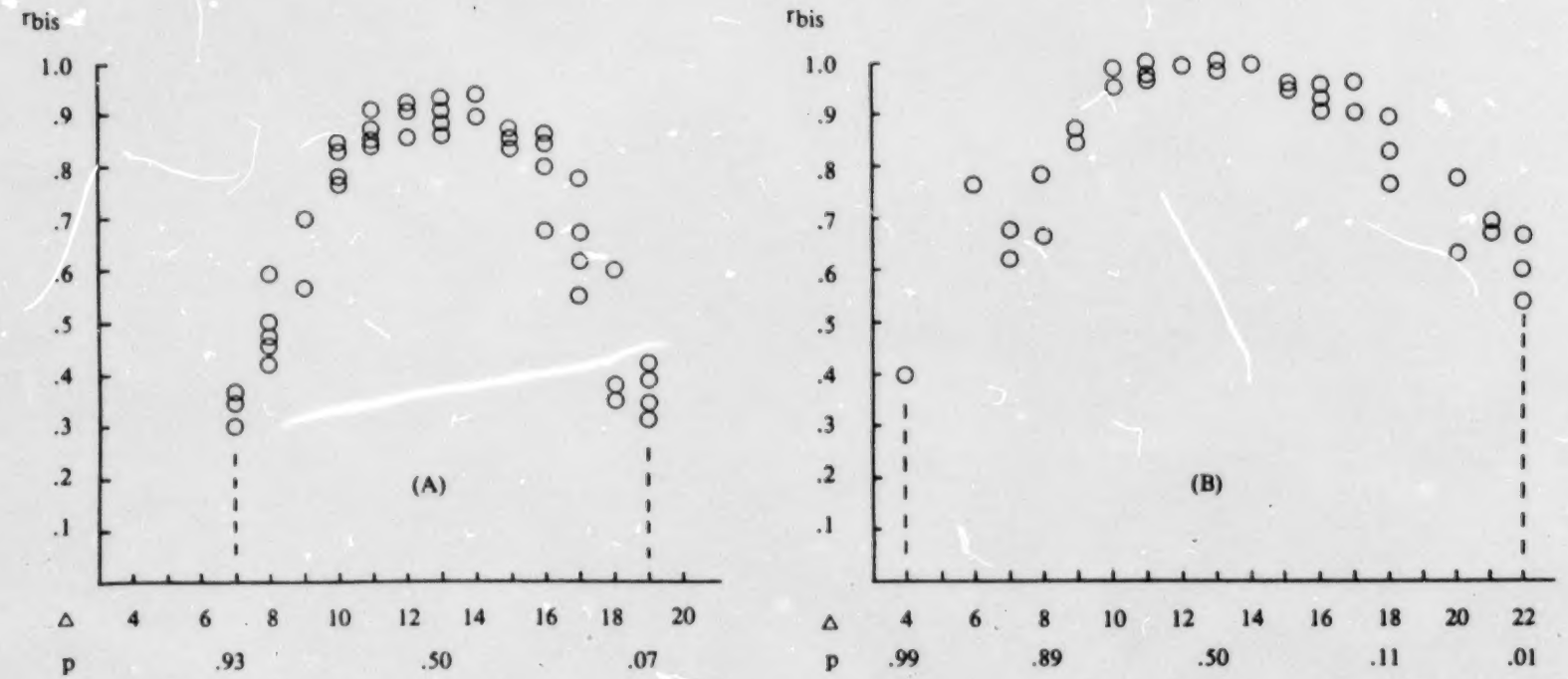

Figure 3. Difficulty and discrimination indices for tests calibrated by classical and Rasch item analysis procedures.

Note: Exact overlap of the data points produces fewer symbols than the number of items contained in each test version.